How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Year: 2023 | *Citations: 1441* | *Authors: Aidan Gilson, Conrad W. Safranek, Thomas Huang*

Abstract

Background Chat Generative Pre-trained Transformer (ChatGPT) is a 175-billion-parameter natural language processing model that can generate conversation-style responses to user input. Objective This study aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams, as well as to analyze responses for user interpretability. Methods We used 2 sets of multiple-choice questions to evaluate ChatGPT's performance, each with questions pertaining to Step 1 and Step 2. The first set was derived from AMBOSS, a commonly used question bank for medical students, which also provides statistics on question difficulty and the performance on an exam relative to the user base. The second set was the National Board of Medical Examiners (NBME) free 120 questions. ChatGPT's performance was compared to 2 other large language models, GPT-3 and InstructGPT. The text output of each ChatGPT response was evaluated across 3 qualitative metrics: logical justification of the answer selected, presence of information internal to the question, and presence of information external to the question. Results Of the 4 data sets, AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2, ChatGPT achieved accuracies of 44% (44/100), 42% (42/100), 64.4% (56/87), and 57.8% (59/102), respectively. ChatGPT outperformed InstructGPT by 8.15% on average across all data sets, and GPT-3 performed similarly to random chance. The model demonstrated a significant decrease in performance as question difficulty increased (P=.01) within the AMBOSS-Step1 data set. We found that logical justification for ChatGPT's answer selection was present in 100% of outputs of the NBME data sets. Internal information to the question was present in 96.8% (183/189) of all questions. The presence of information external to the question was 44.5% and 27% lower for incorrect answers relative to correct answers on the NBME-Free-Step1 (P<.001) and NBME-Free-Step2 (P=.001) data sets, respectively. Conclusions ChatGPT marks a significant improvement in natural language processing models on the tasks of medical question answering. By performing at a greater than 60% threshold on the NBME-Free-Step-1 data set, we show that the model achieves the equivalent of a passing score for a third-year medical student. Additionally, we highlight ChatGPT's capacity to provide logic and informational context across the majority of answers. These facts taken together make a compelling case for the potential applications of ChatGPT as an interactive medical education tool to support learning.